# SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

## DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

## 23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

## UNIT II – DESCRIBING DATA WITH AVERAGES

EMPATHY:

## Describing Data with Averages

- Averages consist of numbers (or words) about which the data are, in some sense, centered. They are often referred to as measures of central tendency.

## Measuring the Central Tendency

- We look at various ways to measure the central tendency of data, include:

Mean, Weighted mean, Trimmed mean, Median, Mode and Midrange.

1. Mean :

- The mean of a data set is the **average of all the data values**. The sample mean x is the point estimator of the population mean μ.

## 2. Median :

- Sum of the values of then observations, Number of observations in the sample
- Sum of the values of the N observations Number of observations in the population
- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, the median is the preferred measure of central location.

## 3. Mode:

- The mode of a data set is the value that occurs with **greatest frequency**.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are **bimodal**.
- If the data have more than two modes, the data are **multimodal**.

## Weighted mean :

Sometimes, each value in a set may be associated with a weight, the weights reflect the significance, importance or **occurrence of frequency** attached to their respective values.

## Trimmed mean:

- A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.
- Even a small number of extreme values can corrupt the mean.
- The trimmed mean is the mean obtained after cutting off values at the high and low extremes.

DEFINE:

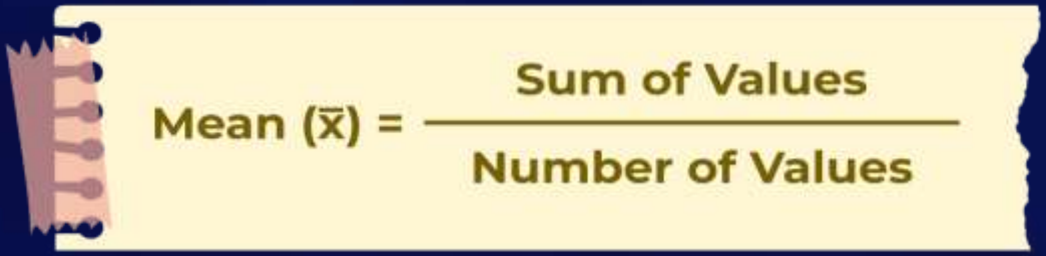
# Mean

Mean  is the sum of all the values in the data set divided by the number of values in the data set. It is also called the Arithmetic Average.

The Mean is denoted as x̄.

Mean Formula

The formula to calculate the mean is: Mean (x̄)  = Σxi/ n



$$\text{Mean } (\bar{x}) = \frac{\text{Sum of Values}}{\text{Number of Values}}$$

**For Example:** Find the mean of data sets 10, 30, 40, 20, and 50.

**Solution:**

Mean of the data 10, 30, 40, 20, 50 is

**Mean = (sum of all values) / (number of values)**

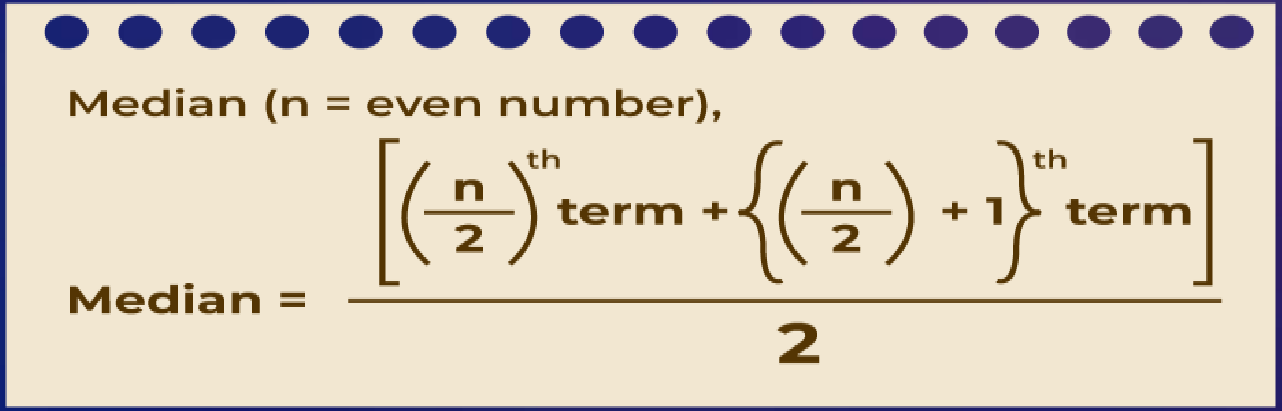
Mean = (10 + 30 + 40 + 20+ 50) / 5 = (150)/5 = 30

IDEATE:

## Median

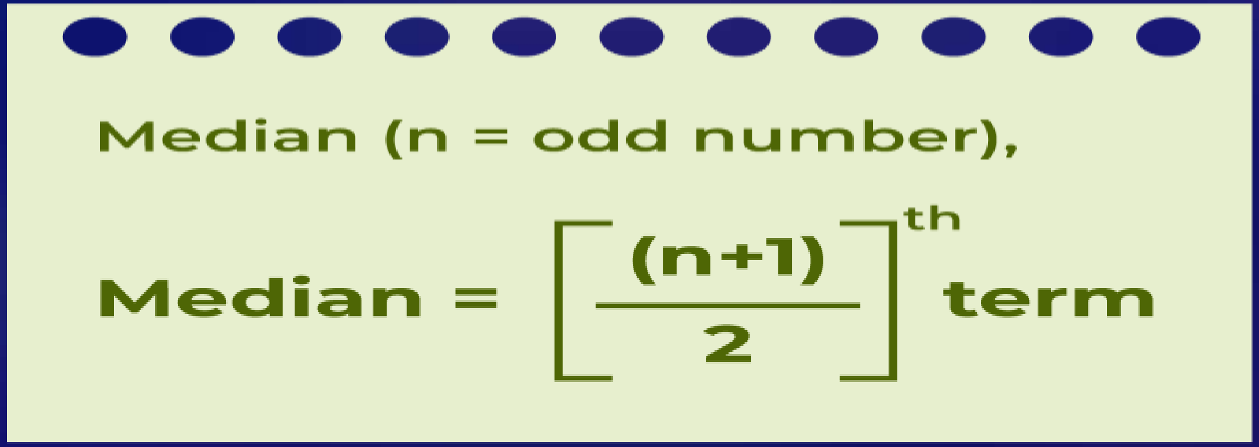
A median is a **middle value for sorted data**. The sorting of the data can be done either in ascending order or descending order. A median divides the data into two halves.

## Median Formula:



Median (n = even number),

$$\text{Median} = \frac{\left[\left(\frac{n}{2}\right)^{th}\text{term} + \left\{\left(\frac{n}{2}\right) + 1\right\}^{th}\text{term}\right]}{2}$$



Median (n = odd number),

$$\text{Median} = \left[\frac{(n+1)}{2}\right]^{th}\text{term}$$

Example: Find the median of the the given data set 30, 40, 10, 20, and 50.

Solution:

Median of the data 30, 40, 10, 20, 50 is,

Step 1: Order the given data in ascending order as:
10, 20, 30, 40, 50

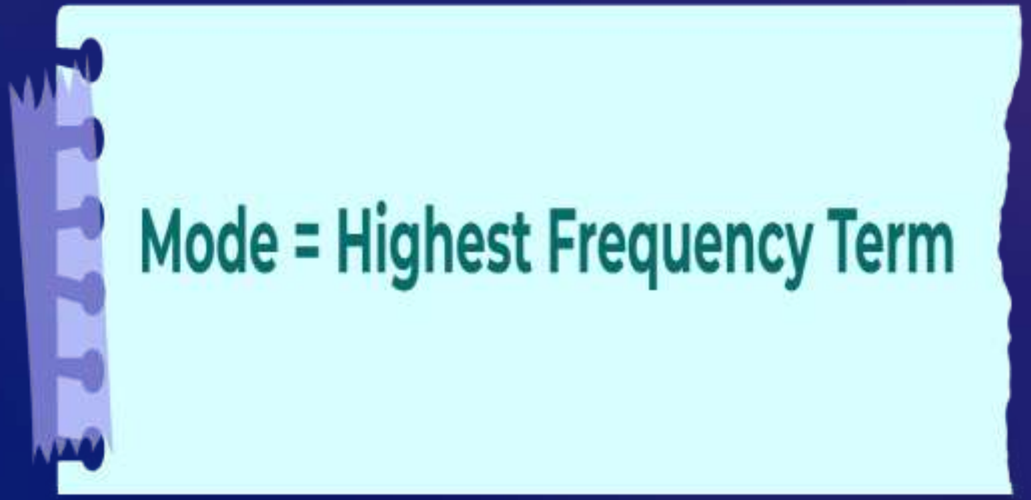
Step 2: Check n (number of terms of data set) is even or odd and find the median of the data with respective 'n' value.

Step 3: Here, n = 5 (odd)
Median = [(n + 1)/2]th term
Median = [(5 + 1)/2]th term

= 30 **Since there are 5 values (an odd number), the median is the middle value**.

# Mode

- A mode is the most frequent value or item of the data set.



Mode = Highest Frequency Term

## TESTING:

Example: Find the mode of the given data set 1, 2, 2, 2, 3, 3, 4, 5.

Solution:

Given set is {1, 2, 2, 2, 3, 3, 4, 5}

As the above data set is arranged in ascending order.

By observing the above data set we can say that,

Using the formula

Mode = Highest Frequency Term

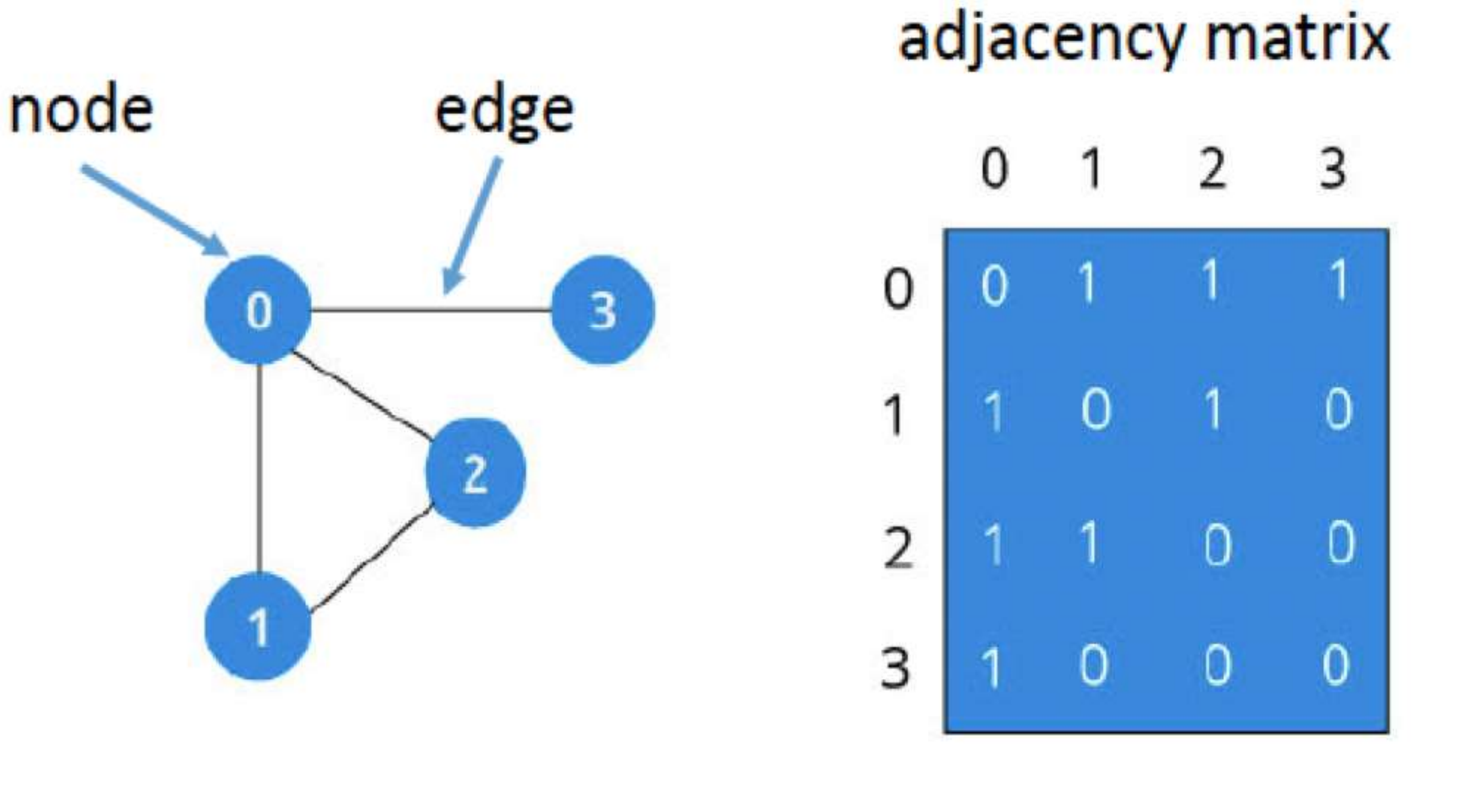
Mode = 2

As, it has highest frequency (3)

# GRAPHS

## Graph Theory Basics

A graph is an ordered pair of G (V, E), where V is the set of Vertices or Nodes and E is the set of Edges or relationships connecting those Nodes such that E ⊆ {(x, y) | x, y ∈ V, and x ≠ y. Refer fig below



adjacency matrix

|   | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 |

## Adjacency Matrix:

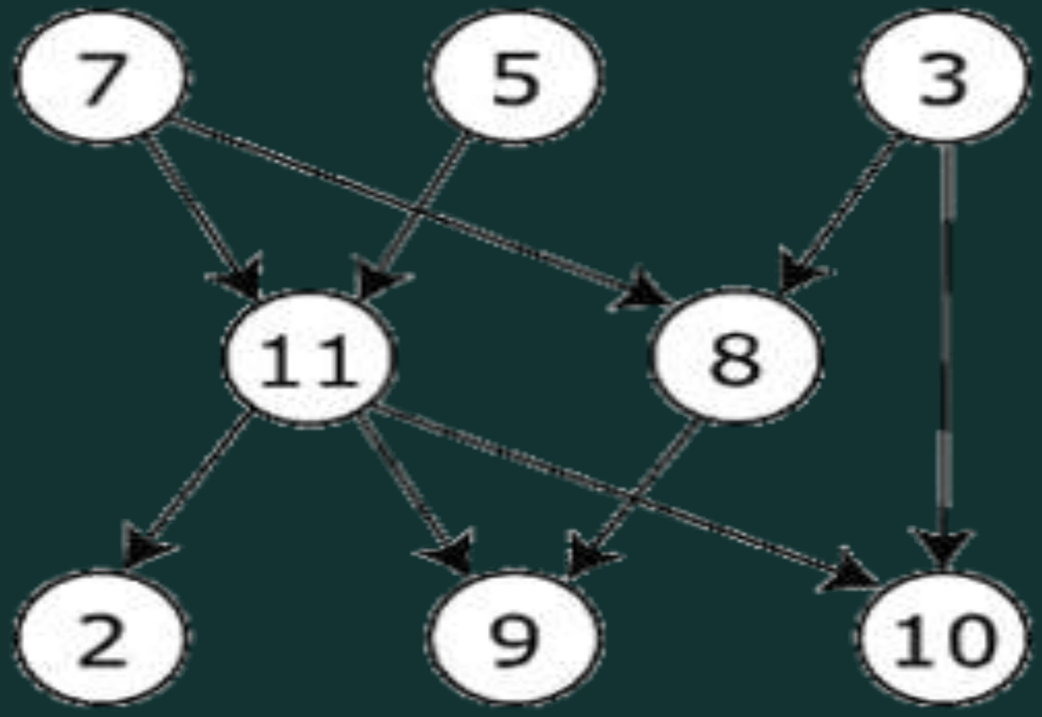
- An adjacency matrix is a square Boolean matrix (comprising of 0’s and 1’s only) representing the graph where the rows and columns are the Nodes of the graph.
- The M(I, J) value of 1 - indicates that Nodes i and j have a direct connection or relationship, and
- M(I, J) value of 0 indicates that Nodes i and j do not have a direct relationship with each other.

## Types of Graphs:

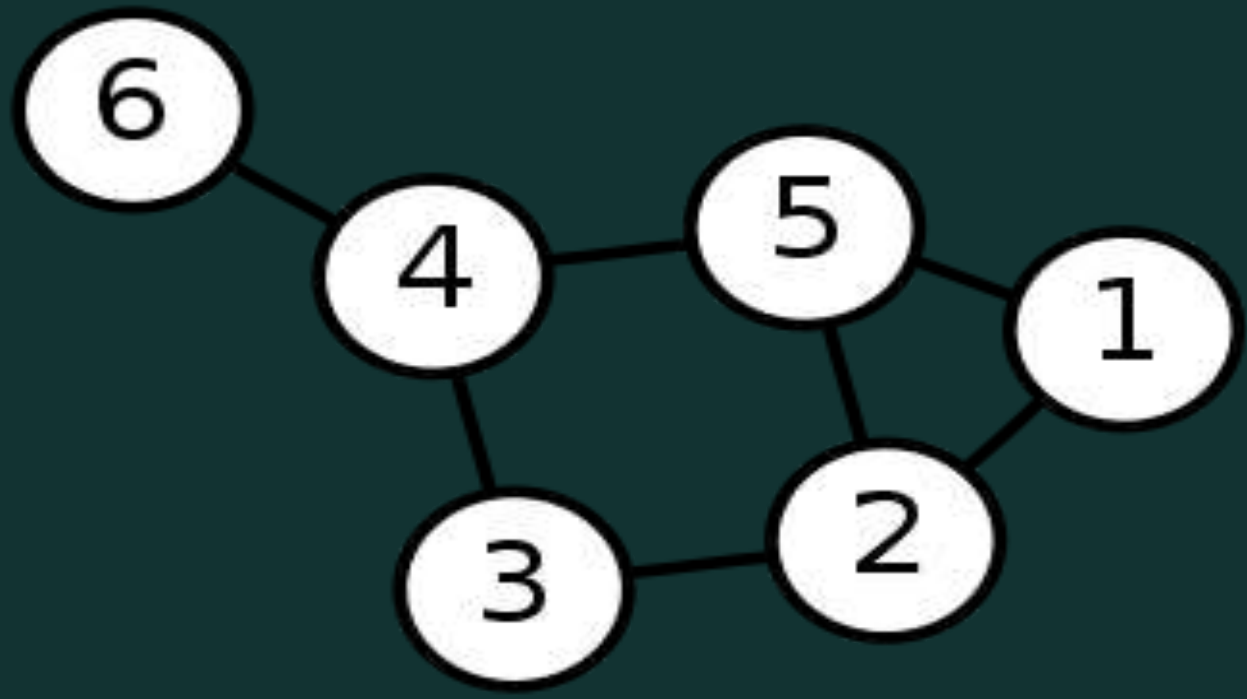
Graphs can be directed or undirected graphs

## Directed and undirected Graph:

In the directed graph, the nodes and links are connected where all the edges are directed from one vertex to another, whereas an undirected graph is a set of nodes and links between the nodes.



Directed graph



Undirected graph

## Cyclic Graph and loops:

### Loops:

- In graph theory, a loop or a self-loop is a node that connects a vertex to itself.

### Cyclic and Acyclic graphs:

- Whenever in a graph, a few vertices are attached in a closed chain of relations, then the graph is said to have a cycle.
- A graph with at least one such cycle is called the cyclic graph, and the graph with zero cycles is an acyclic graph. In the above-undirected graph, the nodes 2,3,4,5 form a cycle

## PROTOTYPING:

Graph Operations

Before performing some operations on the graph, we created above, let us look at a few more concepts from graph theory that are used in graph data science

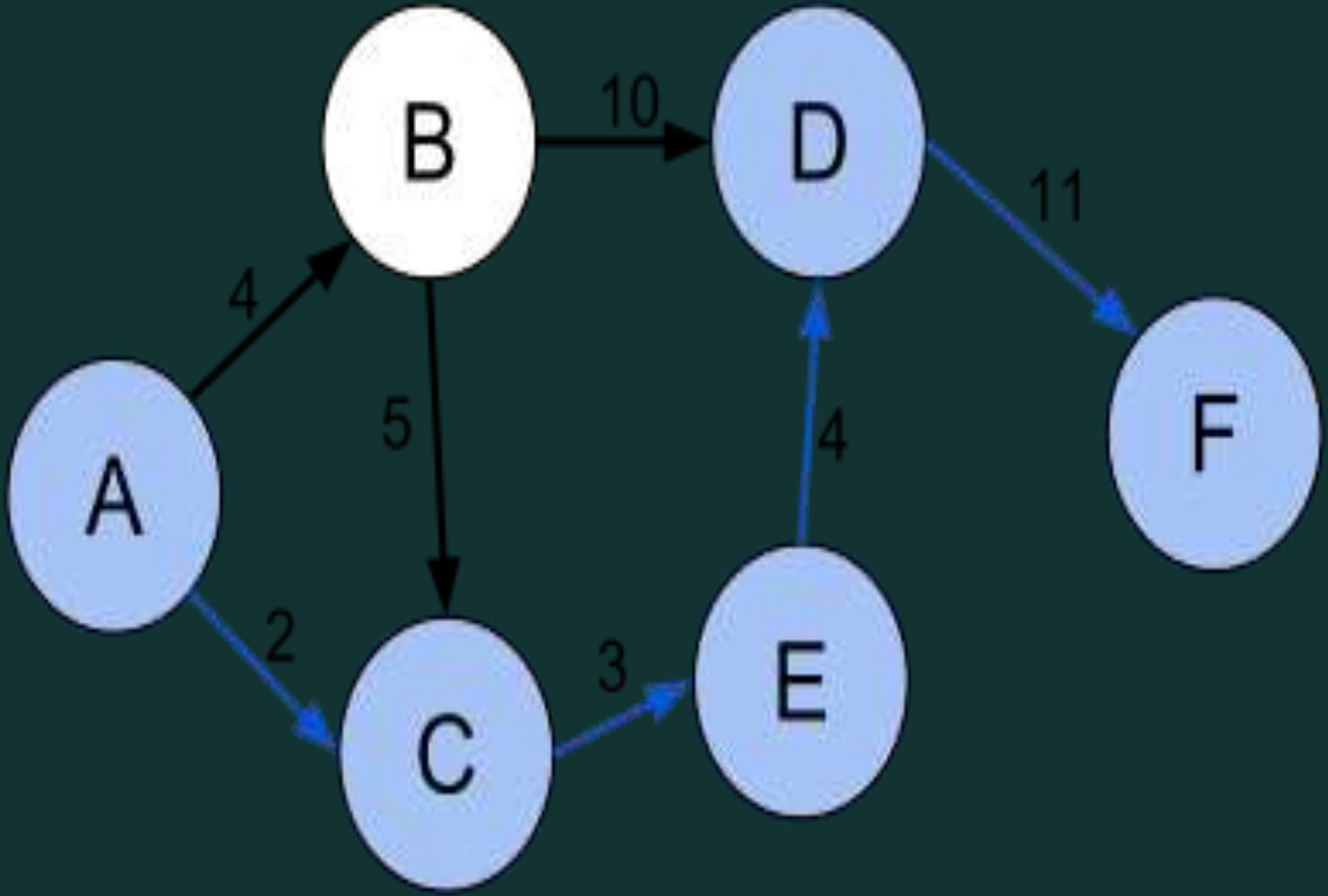
Path in Graph:

- A path in a graph is a finite or infinite sequence of edges connecting distinct vertices between two nodes.
- In a directed graph, the node order cannot be reversed.

Shortest Path between any two nodes in a graph:

- The path between two nodes that has the **least sum of the weights** of its constituent links is the shortest path.

- For example, in below figure shortest path (A, C, E, D, F) between vertices A and F in the weighted directed graph is 2+3+4+11 = 20



B
10
D
11
4
5
4
F
A
2
3
E
C